DOCUMENT RESUME

ED 091 421 TM 003 635

AUTHOR Popham, W. James
TITLE Technical Travails of Developing Criterion-Referenced Tests.
PUB DATE [Apr 74]
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; Educational Objectives; Evaluation Needs; *Measurement Techniques; *Test Construction; Testing Problems
IDENTIFIERS *Instructional Objectives Exchange; IOX

ABSTRACT

Since mid-1971 the Instructional Objectives Exchange has been engaged in a major effort to develop and disseminate criterion-referenced tests in the fields of reading, mathematics, language arts, and social studies. This paper isolates the chief technical decision-alternatives faced in this project, such as: (1) the optimal number of tests to produce, (2) choosing a well defined behavior domain, (3) devising a homogeneous domain, and (4) the high cost of preparing criterion-referenced tests. The author describes the rationale for each decision, then appraises the adequacy of the decisions on the basis of empirical results. (Author/MLP)

TECHNICAL TRAVAILS OF DEVELOPING CRITERION-REFERENCED TESTS*

W. James Popham
University of California, Los Angeles

Despite the fact that numerous educational practitioners are becoming disenchanted with norm-referenced achievement measures of evaluation purposes, until recently there have been few reports of large scale efforts to construct criterion-referenced tests which might serve as alternatives. Perhaps the paucity of major projects designed to develop criterion-referenced tests provides a partial explanation for our extremely limited progress in devising the technical procedures needed to construct criterion-referenced tests. Maybe we have not yet acquired a sufficient experiential base to permit the invention of respectable counterparts to our well honed procedures for establishing the validity, reliability, and item adequacy of norm-referenced tests.

On the other hand, perhaps the absence of many major efforts to devise criterion-referenced tests is partially attributable to a recognition of today's primitive state of criterion-referenced measurement methodology, and a concomitant trepidation about plunging into a major engineering effort where the toolshed is so barren.

But whether cause or effect, measurement personnel obviously can learn much from those few large scale projects which have produced criterion-referenced tests. Recent announcements of the availability of criterion-referenced measures (in reading and mathematics) by such firms as CTB/McGraw Hill and Educational Development Corporation attest to the likelihood that a number of test developers will soon have experiences to share and, hopefully, technical advances to trade.

Certainly, a number of state and local education agencies have set out to enter the criterion-referenced measurement game in a major way. To mention but a few, formidable test development activity is underway in such states as New Jersey, Michigan, Oregon, Wisconsin, and California plus a number of large metropolitan school districts throughout the nation. But we need to share our experiences, both successes and failures, in attempting to cope with the technical problems presented when one undertakes the development of a large number of criterion-referenced measures. The paper, consistent with an experience-sharing mission, describes practical dilemmas and efforts to provide technical solutions during an ongoing test development project conducted at the Instructional Objectives Exchange (IOX), a Los Angeles-based non-profit educational corporation. The paper could aptly be subtitled: Confessions of a Criterion-Referenced Test Developer, since it will recount, as candidly as possible, the dubious decisions, outright errors, and glittering insights (both of them) made during the IOX criterion-referenced test development enterprise.

*A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, April 16-18, 1974.

Background

Perhaps better known as a depository of measurable instructional objectives, since to serve as such a depository was the reason for its establishment in 1968, the Instructional Objectives Exchange embarked on a major test development effort in the summer of 1971. In an attempt to provide test instruments which would more readily incline educators to organize their instruction and evaluation activities around measurable learner outcomes, the IOX staff initiated a project to develop criterion-referenced measures in all major subject fields taught in public schools. During the first two years of the project's operation, criterion-referenced tests in reading and mathematics were produced. During the third year, tests in language arts and U.S. Government were developed. Currently, more advanced tests in mathematics and language arts are being prepared. At this point we have almost three years' worth of experience[1] in producing these tests, and it is time to engage in a little stock-taking, soul-searching, and agonizing reappraisals. This paper will focus on the major problem areas which have thus far been isolated.

An Optimal Number of Tests

As will soon become apparent, one of the more pervasive problems facing those who would work with criterion-referenced tests stems from our almost complete naivete regarding what level of content or skill generality should tests incorporate in order to function optimally for purposes of educational evaluation. Although curriculum and measurement specialists have reminded us for decades that this represents a critical, unresolved problem, we are not substantially closer to a solution today than we were at the start of this century.

The first place the generality-level difficulty manifests itself is in connection with determining how many criterion-referenced tests to produce. Once the IOX staff had decided to develop a set of criterion-referenced measures which would be of value to educators throughout the nation, the first problem we had to deal with was, "How many tests?"

Educators learned an important lesson from their 1960's preoccupation with behavioral objectives, namely, that increasing the specificity with which an objective is stated does not necessarily increase the educational utility of that objective. The difficulty that arises when we try to make our objectives too specific is that we end up with literally thousands of such super-specific objectives. As an objective becomes more specific, its scope is generally reduced, thereby obliging the educator using the objective to keep track of more objectives than is feasible.

The situation with respect to criterion-referenced test construction is analogous. We could produce several hundred tests of the various

[1] After the first year of the project's existence, a technical paper was produced delineating the various technical procedures employed during the initial year of the activity: Popham, W. James, Procedural Guidelines, Developing IOX Objectives-Based Tests, Instructional Objectives Exchange, Los Angeles, California, August, 1972.

skills required to assess an individual's mastery of reading competencies. But, by trying to test all of the isolatable skills required in reading, we would have created a battery of tests so awesome in bulk, let alone conceptual complexity, that few sane teachers would use the tests. And odds are that if teachers did use such tests for any extended period, all remnants of sanity would vanish.

Thus, we knew we had to go with fewer tests, but how many fewer was up for grabs. We tried to approach the problem by thinking about the number of outcomes (as reflected by student test performances) that a typical teacher could realistically monitor. We concluded, on the basis of shared personal experiences, that somewhere between a half-dozen and a dozen[2] outcomes per course per year could be meaningfully dealt with by most teachers. This meant, for example, if IOX was developing a set of reading tests to be used in grades K, 1, 2, and 3, we would try to develop somewhere between 24 (six tests times four grade levels) and 48 (twelve tests times four grade levels) instruments.

But this mechanism for reaching a decision regarding numbers of tests (personal estimates of teachers' capacity to process test information) certainly induces little confidence. It represents a primitive form of pooled guessing to answer a question which is amenable to an empirical solution. There is no reason to estimate how many tests teachers can meaningfully monitor when we can carry out studies which will demonstrate what teachers' preferences truly are. A variety of straightforward investigations could be concocted, for example, in which illustrative test results of varying degrees of complexity, content coverage, etc. were presented to teachers for their reactions. Teachers could register preferences for different types of test data, different numbers of tests, etc. It would also be possible, though more demanding, to follow up teachers who had been presented with differential test information to see which form of the information led to more meaningful teacher utilization of data.

The point being made, of course, is that as long as we are forced to rely on intuitive estimates of how many tests teachers can keep track of, our estimates are apt to be some distance from reality. We need to embark on a programmatic research effort to understand more fully the level-of-generality problem as it applies to determining the optimal number of tests.

In the meanwhile, we can eschew the extremes regarding test numbers, i.e., two or three tests are too few; two or three hundred tests are too many. The limits of that tolerance band are too wide, unfortunately, to be of much help to the criterion-referenced test constructor. Those individuals and agencies who do not deal with this problem prior to their decision to generate criterion-referenced measures, are failing to confront an issue that may well render their development effort ineffectual.

[2] Several IOX staff members possessed previous experience in bakery and donut shops, leading to a proclivity to think in terms of dozen or $\underline{\text{dozen}}$.
2

Choosing a Domain

When the IOX staff began to create criterion-referenced tests we had to step back several paces and remind ourselves why it was that norm-referenced measures were not satisfactory for evaluative purposes. Although there are a number of reasons why this is so, the chief deficit of norm-referenced measures is that they fail to provide a satisfactory description of what examinees can or cannot do. Since norm-referenced measures yield scores interpretable according to the examinee's relative standing with respect to a norm group, it is often difficult to obtain a clear idea of what the dimension is on which examinees are differing. Criterion-referenced tests, on the other hand, are used to ascertain an individual's status with respect to a well defined behavior domain, that is, class or set of behaviors. It is this well defined behavior domain, in fact, which constitutes the "criterion" to which examinee performance is referenced. To the extent that a criterion-referenced test fails to provide an explicit description of what it is that the examinee can or cannot do, it offers few advantages, at least for purposes of evaluation, over a norm-referenced measure.

In view of the need to provide measures which yielded better descriptions of learner performance, the IOX staff drew heavily on the work of Wells Hively and his associates[3] who had been working since the mid-sixties to devise ways of delimiting classes of learner behaviors for purposes of curriculum development and test design. The item form approach used by Hively provided the IOX project with its chief model, although, as Hively and his cohorts used them, they were typically too complicated for sustained use, either by our item writers or by the public school educators who would be relying on them for interpretation purposes.

We saw two criteria as important in divising our domain descriptions, namely, clarity and brevity. These two criteria are typically in conflict. We tried to produce domain descriptions which were detailed enough to delimit the class of learner behaviors to be measured, but short enough to be used for interpretive purposes by busy educators. The task was a difficult one.

Because of the IOX association with instructional objectives, we referred to our domain descriptions as amplified objectives, since we essentially elaborated on a simple statement of instructional objective in order to produce a domain description.

Thus, we settled the question of what our domain description would more or less look like, that is, a few paragraphs which attempted to describe the nature of (1) the stimuli presented to the examinee, (2) the response options available, and (3) the criteria for judging the examinee's response. But as we set out to delineate the domains for our tests, we

[3] Hively, Wells; Maxwell, Graham; Rabehl, George; Sension, Donald, and Lundin, Stephen. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemost Project, Center for the Study of Evaluation, UCLA, 1973.

ran into an unexpected problem. For any general skill that we tried to operationalize via a more explicit statement of a learner behavior domain, we found that we had several clearly competitive domains. The trick was to decide on the best representative from the contending domains.

Let's illustrate this problem a bit more tangibly. Suppose we wanted to assess whether a student could satisfactorily solve word problems involving elementary multiplication operations. Now one very clear difference in the kinds of domains we might choose would hinge on whether our story problems called for a constructed learner response, as when the child is called on to write out answers to problems, or whether the problems required selected responses, as when the student chooses an answer from several alternatives. Other variations in the domain might involve the nature of the key ingredients in the story problems, such as the exact kinds of eligible problem types that could be presented to the learner, that is, what would be missing, what kinds of distractor data would be included in the story problem stem, and so on.

Anyone who believes that, when setting out to build criterion-referenced measures, it will be obvious which domain of behaviors should constitute the criterion (to which test items will be referenced) is in for a disappointment. The alternatives will be numerous. The selection decisions will be difficult.

At IOX we tried to approach the task somewhat rationally by setting forth criteria to guide our developers. The following six considerations were to be used in deciding on domains:

1. General Acceptance. Is the behavior domain considered important by teachers, subject matter specialists, the public?
2. Transferability Within the Domain. Since the learner behavior measured by the test is highly specific, will that behavior, when mastered by the learner, be likely to transfer to similar skills (other domains) within that general class of behavior desired?
3. Transferability Outside the Domain. Will the domain, once mastered, be likely to transfer to learner behaviors required in rather different types of behavior domains?
4. Terminality. If the organization of learner behavior is hierarchical, will the domain selected tend to be terminal rather than en route?
5. Amenability to Instruction. Will the domain measure a learner skill that can be taught, rather than a native trait relatively immune to instruction?
6. Ease of Scorability. Other factors being equal, will the domain selected yield learner responses which can be easily scored, not necessarily objectively scored, by those educators using the tests?

But even though our staff has attempted to use these criteria in deciding which domains to select (create), we have not subjected our decisions to any kind of empirical verification. The most important question we should probably ask is, "How generalizable will the skill be that is reflected by the learner's mastery of this particular domain?" It would appear that this question is amenable to an empirical answer and that the methodology for treating this issue will resemble but not coincide with those tactics ordinarily employed when one attempts to secure construct validity data. This question is treated at greater length elsewhere,[4] but it is a critical concern for those devising criterion-referenced measures. Technical procedures for dealing with this problem are desperately needed.

Domain Homogeneity

A third technical problem area we have encountered in our IOX test development operations arises from the same difficulty we encountered when deciding on how many tests to prepare, that is, the level of generality question. While it might be possible to lump both problems under the handy rubric of generality-level, the problem takes a different shape when faced by those who must generate domain definitions. Let's try to clarify the issues.

In the abstract, it would be desirable to devise criterion-referenced test domains so that they permit the generation of a pool of items which would not only *appear* to be homogeneous but would also, on the basis of empirical data, *function* in a homogeneous manner. Yet, in order to create domain definitions which possess these qualities, we found that we were isolating only a very discrete type of learner behavior. For example, suppose we are dealing with a domain which taps the learner's ability to divide words into their constituent syllables. Now when one analyzes the various kinds of syllabication tasks which might be included in a domain description, even assuming we had decided how the learner would identify syllables, there are several *discernibly different* sorts of tasks which might be included. Most of these would be contingent on the kinds of words with which the learner will be presented. There are, for instance, simple two-syllable words in which the syllable-division task is straightforward, e.g., Oxford. Then there are two-syllable words where the novice could not quickly discern whether a middle consonant belongs with the first or second syllable, e.g., color. There are a number of other classes of words with varying numbers of syllables and varying rules for syllabication. Now the technical decision facing the domain writer is whether to include these different sorts of syllabication tasks in a single domain, more or less randomly, or to deliberately try to include a preset number of such tasks.

Some would recommend that we adopt a *diagnostic strategy* by including a number of the different subskills in a single domain in order to detect which particular subskills the learner can perform. But when you do this, of course, you're obligated to include a reasonable number of the learner's prowess with respect to each subskill. This leads to lengthy tests, too lengthy sometimes to be practical.

[4] Popham, W. James, *Systematic Educational Evaluation*, manuscript in preparation.

Others recommend that we adopt a medley strategy, tossing in all kinds of different item types with the anticipation that some sort of global estimate of learner skill will emerge. The problem with this approach, of course, is that with only a modest number of items per domain, a few items representing a particular subskill can inequitably influence the learner's performance on the domain, and we will be unaware of what subskills are influencing a score on the total domain.

A third possibility is to opt for a terminal skill strategy where we isolate only the learner skill which is the most terminal, then measure that skill. Typically, of course, this leads to measuring the most difficult of those skills which might be included in the domain. While this approach characteristically yields a homogeneous domain, it is obviously of little value for any type of diagnostic purpose.

I would like to be able to report that the IOX staff, having considered these alternatives rationally, has completed a series of empirical investigations which demonstrate conclusively what the proper approach should be. I would also like to be ten years younger, twenty years wiser, and forty times wealthier. The distressing truth is, however, that in the IOX project we have adopted a stance of unflinching vascillation. We really don't know how to proceed with respect to this problem.

Costs and Conscience

In 1970 IOX began to produce criterion-referenced tests because, frankly, we were tired of waiting around for major test publishers to do it. We were well aware of the deficiencies of norm-referenced measures for purposes of educational evaluation, but could find few commercially available alternatives to such tests. We were unwilling to continue to tell educators that, while norm-referenced tests were to be avoided, criterion-referenced counterparts would be available only when the measurement princes in Princeton, New Jersey (and elsewhere) got off their duffs.

So, we went into the criterion-referenced test development business. What an expensive business it is! Although any kind of test preparation is costly, we had no idea how much personnel time would be tied up in devising domain descriptions, generating hopefully congruent items and monitoring item-domain congruency. Maybe the measurement princes weren't so dumb.

The choice facing us quickly became apparent. Either we could proceed with our developmental activities, having recognized that our self-support financial base was far too modest to permit the kind of quality we wanted, or we could fold up our development tents with the hope that other, better funded organizations would get around to producing decent criterion-referenced measures. It was a real choice point for us. It will be a real choice point for most agencies that embark on a criterion-referenced test development effort.

As might be inferred from the fact that our test development efforts are still underway, the IOX staff decided to go ahead with their test creation activities. In retrospect, even though we suffer periodic conscience pangs because our tests and domain descriptions aren't as flawless as we wish, I am pleased we decided to continue. A number of the commercially distributed criterion-referenced tests now making their way to the educational market suffer from inadequate or non-existent domain descriptions as well as an inadequate number of items per domain (IOX tests have either five or ten items for each domain). The IOX tests at least provide an alternative.

The more fundamental question emerging from this problem, howeve is whether a commercial or nonprofit test development agency, *sans substantial external subsidies*, can engage in the development of truly high quality criterion-referenced tests? I believe, although there will be occasional exceptions to the rule, that the answer is *no*.

A Technical Wasteland

This leads to a related, and final, observation. There is a growing force within the educational community to turn from norm-referenced measures with their technical deficits (for evaluation purposes) and to replace them with more sensitive criterion-referenced assessment devices. That's just fine.

At least it would be just fine if we had any assurance that the criterion-referenced measures which will be produced in the next few years will be capable of doing the jobs educators want accomplished. Unfortunately, because the technological support base for criterion-referenced measurement development is so perilously weak at present, our predictions regarding the quality of future criterion-referenced tests must be gloomy.

What we need, and now, is a well financed, governmentally-initiated project to expand our weak technological base in this crucial measurement area. There are, at this writing, no major efforts underway to sharpen the technical tools needed to produce better criterion-referenced measures. Because of the pivotal role to be played by such measures in all sorts of evaluation and accountability programs, this situation is intolerable.

We must, without delay, muster whatever clout we have in order to encourage the National Institute of Education or some comparable agency to foster the kind of technical-support activities which will lead to a reduction of the technical travails associated with developing criterion-referenced tests. The decisions which these tests will influence are too important to treat the tests with clumsy crowbars. We need scalpels.

b13w